下载APP

22 | 采集数据：用好一手数据和二手数据

2021-09-17 郭炜

《数据分析思维课》 课程介绍 >

讲述：郭炜

时长 17:00 大小 15.59M

数据给你一双看透本质的眼睛，这里是《数据分析思维课》，我是郭炜。

在上面一节课里，我们讲了如何确定我们要分析的数据问题，为整个数据分析的过程指明了方向。整个数据分析过程就像规划设计一个大的数据地图（就像玩 Simcity），先有了大的城市框架，再去规划每一个街区。

在规划这一步，测绘、逐步细化迭代非常重要。所以今天我就来给你讲讲测绘——也就是我们的数据采集。

我们进行问题决策的时候，如果没有数据采集，就会陷入经验主义，通过拍脑袋来进行决策，这不是数据分析思维的主张的方向。

现在我们有了方向和问题列表，那么数据会从哪里来呢？

我们在收集数据的时候，数据的来源会分为两大类：一手数据和二手数据。根据这些数据，我们会进行数据探索并产生一些衍生数据，最终为我们下一节课的数据分析思路组织与撰写提供弹药。

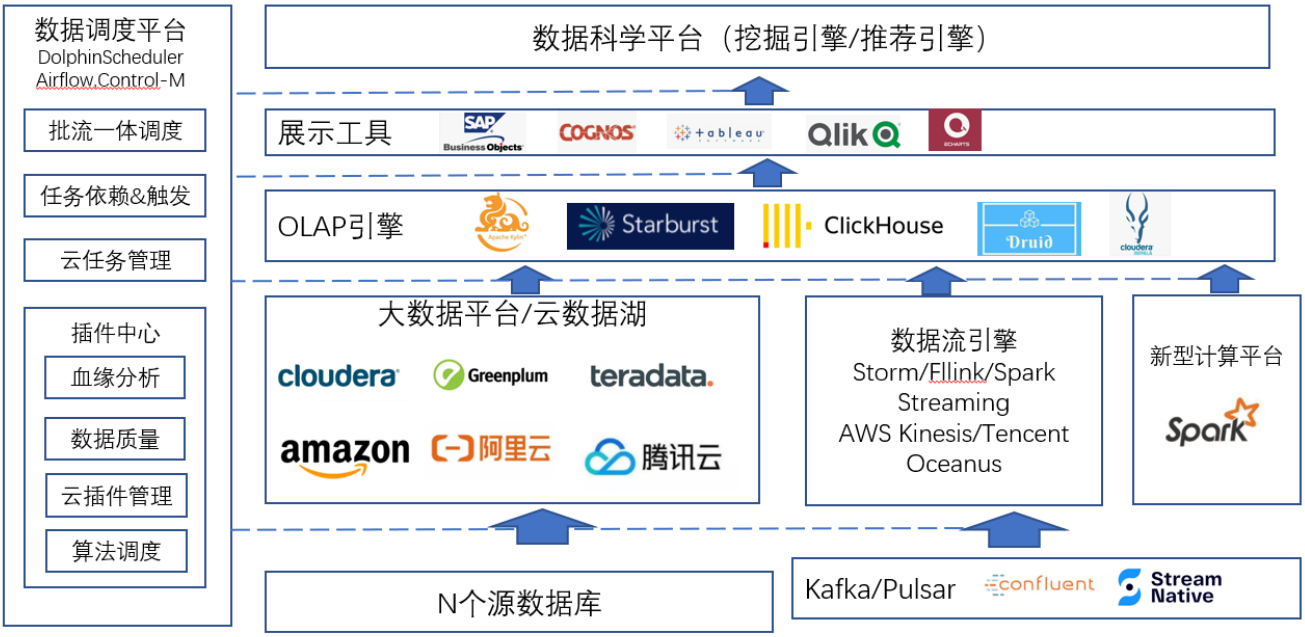
采集数据类型

我们先来看采集数据中的第一类数据来源：一手数据。一手数据主要来自企业内部的大数据平台、数据仓库以及相关系统，还有部分数据来自用户访谈和调研问卷以及内部沉淀的历史文档。

一手数据的特点就是数据可控，也正是因为数据都掌握在自己企业手中，理论上只要付出成本，我们可以拿到所有的想要的数据。也就是说，我们可以通过数据采集、建立相关流程业务系统来进行录入，或者开展大规模的用户访谈以及调研问卷去采集到我们想要的数据。这是一个企业数字资产积累的过程，现在很多企业都已经在关键业务流程上实现了数字化升级和转型。不过在数字化转型升级过程当中，有三点我觉得尤其需要注意。

第一，数字化升级转型应该先从核心或者创新业务流程开始。

这也就是前面一节课说的和收入、支出直接相关的系统优先升级。例如现在传统企业做互联网升级，所有相关的互联网用户画像、用户行为采集、广告投放以及相关的财务数据系统就要优先建立。同样，智能制造企业会优先建立物联网数据采集、物联网大数据平台、供应链决策支撑系统等。现在数据技术手段已经非常发达，下面我给你总结了当前最新的企业大数据分析架构，你可以参考一下。



第二，数据的采集和计算一定要从最明细的数据开始。

不要使用企业内部的二手数据，也就是不要拿加工后的数据再进行加工分析。因为但凡这样做了，数据质量和数据治理这部分的问题往往会花费我们特别多的时间。

其实获得一份精美数据的效率是一个企业能否成为典型数据驱动型企业的明显标志，你可以看到像阿里、腾讯这种典型的一线互联网大厂，都可以提供给数据分析和运营部门相关权限的所有明细数据。数据分析师都是可以在分钟级甚至秒级来给出相关数据统计分析答案，而不是通过层层数据二次加工获得。数据分析师可以直接自己定义计算口径，针对明细数据进行数据探索，这是数据分析的基础要求。

第三，在做数据分析的时候，数据质量的要求要大于数据量的要求。不要迷信于大数据的量大，而是要关注高质量的小型数据，往往几个 Pb 的大数据会比小数据更难获得知识（你可以去复习一下数据抽样这节课的内容）。

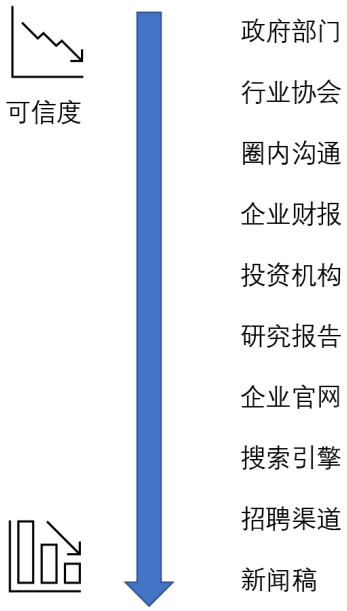
我们接着来看采集数据中的第二类数据来源：二手数据。二手数据主要是来自行业内的数据，它们不是自己企业内部产生的。

一般二手数据用于让我们看到行业内的竞争对手或者整体行业的趋势。例如，如果你是互联网行业的，你可以看到互联网里面用户活跃到底怎样，留存度如何？或者某些广告投放的转化率以及均值如何。这些数据可以帮助我们分析自己公司在行业里整体的水平，从而提高我们自己的目标值。

一般二手数据来自政府部门的报告、行业协会、企业财报、投资机构还有企业官网和一些新闻稿，同时也会来自圈内的沟通或者行业内专业咨询公司出的专业分析报告。我把自己常用的一些网站和信息来源渠道列在了附录里，你如果想找一些二手的渠道的数据，可以到附录去进行搜索。

你要特别注意鉴别二手数据当中的可信度，因为很多企业为了扩大市场影响力，经常会使用“数据技巧”来修饰数据，这样可能会出现因果倒置或者前面讲到的各种数据问题。比如我们从新闻稿里看到某企业收入复购率提升了 100%，如果我们不看细节会觉得这个提升很高，但是如果仔细看文章和企业年报的数据，你会发现它的复购增长率是从第一年的 10% 上涨至第二年的 20%，尽管确实上涨了 100%，但整体上对于企业来讲，这个复购率还是非常低的。

这也是为什么我们在企业内部使用数据的时候，我要求不要使用二次加工的数据的原因。为了鉴别这些数据的真伪，你可能需要熟练掌握数据分析基础这一章里所有的数据工具和方法，才能不被数据所欺骗。如果我们根据数据来源的可信度进行排序，会得到下面这样一个图。



数据探索

经过上面这一步，现在你获得了大量的原始数据，但这对于你解决你最终的问题依旧不够，我们要进行数据探索，将前面采集到的多种数据进行横向纵向的深度挖掘，才能发现

其中的一些原因和知识。在这里我给你介绍常用的三种拓展方法：趋势分析法、快照扩展法和衍生指标法。

趋势分析法

趋势延伸法就是我们找到某一个类型的数据之后，捕捉这个数据一个时间段以内的变化。通过这些数据变化，我们去知道曾经有哪些变化、对结果数据会有哪些影响，这样可以找到其中关键的问题和原因。

这个时候我们经常会用前面介绍的折线图、散点图和回归来分析趋势并确定离群点。我们要尤为关注离群点，因为这些离群点发生的原因往往就是解决问题的答案。

看整个趋势的时候，我们要注意到那些呈指数分布增长的数据，他们往往是对我们非常有意义的。而对于比较平直的曲线来说，我们需要关注整体数据的波动情况，也就是看离散系数是不是很大，因为这代表着这个业务的稳定性。

如果上面说的这些你有些记不清了，建议你带着现在的问题去复习一下我们第一章数据分析基础里的内容。

快照扩展法

快照扩展法是截取某个时点的情况，然后通过下钻的方式来扩展这个指标的分布情况。我们会看在这个时点里面我们各部分对于整体的占比和影响程度。

这里我们会经常用到曾经讲过的直方图、散点图、聚类分类和数据分布当中的方法，来看各个细分渠道、细分部门分布情况，从而找到我们重点要分析的部门、渠道或一些重点的原因。这样做其实是为了明确我们分析范围的目的，把所有的数据信息全都放到一起就像一笔糊涂账，一个有效的数据信息也无法拿到。

衍生指标法

如果用上面两种分析方法还没有找到其中的原因，我们可以进一步进行数据的加工，制造出一些衍生指标来拨开迷雾，这也就是衍生指标法。

优秀的衍生指标就像几何当中的辅助线一样，会帮助我们看到更有意义的数据。例如当我们看到售卖产品的数量和我们的广告投入几乎无关，那么我们应该意识到，只是看收入和投入的这个表面关系，很难做出恰当的评价。

那么要想衡量这样的数据，我们就会建立新的指数——用户忠诚度指数。我们通过这个指数去衡量我们获客之后，这个客户会不会再次购买我们的产品。关于建立衍生指标的方法，你可以去复习一下第 10 节课，我们要学会像建立智商指数和上证指数一样，自己建立新的衍生指标去定义和分析数据里面的相关内容。

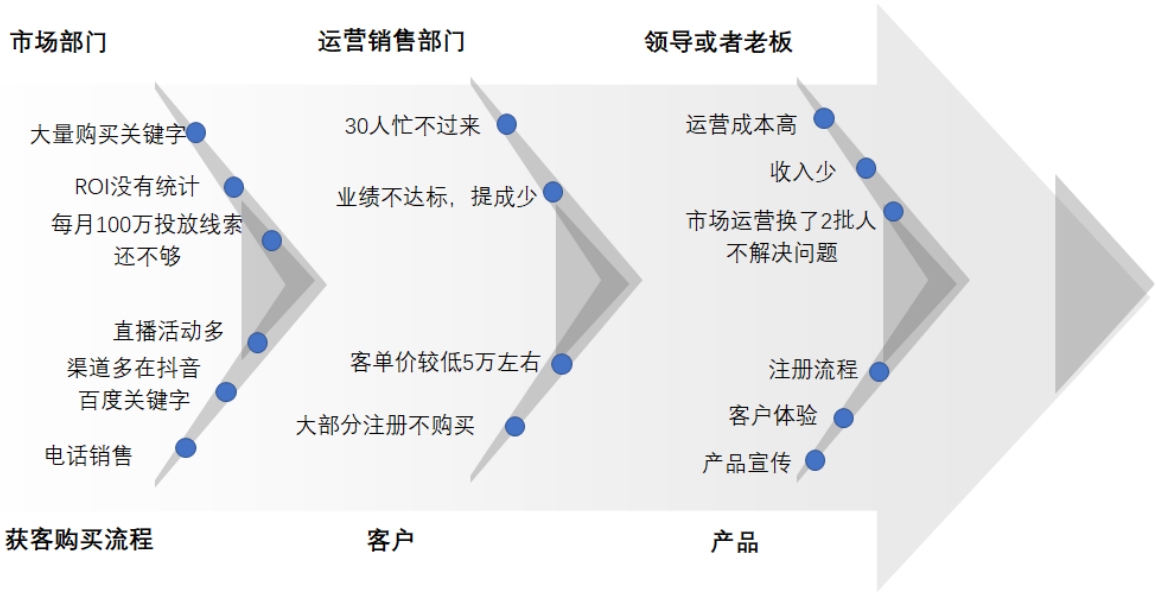
在进行数据探索的时候，有三个点你需要重点注意一下。

关注数据质量的把控。例如我们在进行新冠统计的时候，往往你会发现统计死亡率要比统计得病率更加准确。

注意避免辛普森悖论。这就要求我们在看快照扩展法状态值数据的时候，尽量细分领域和时间。

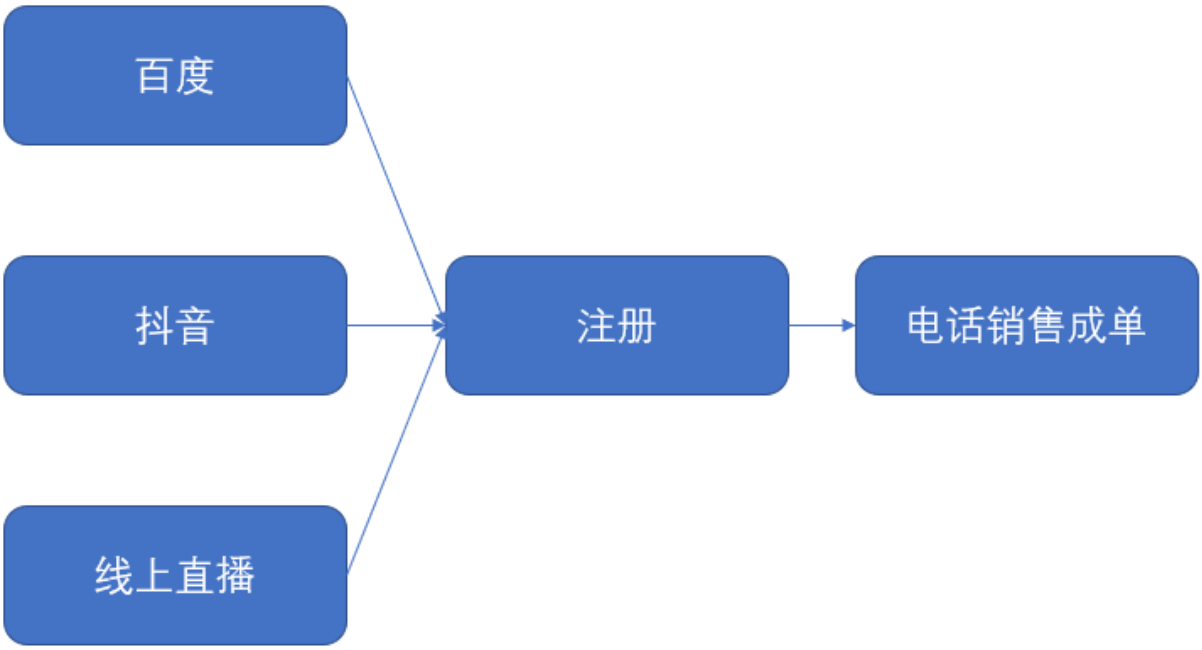
注意避免因果倒置。例如你在整体沿着大思路进行分析的时候，看到了看广告用户的转化率和没看到广告的用户的转化率，你要能够客观去做衡量。

具体示例



在上节课里，我们其实已经梳理了所在企业可能的一些数据分析方向和问题。这里我们选择其中一个方向进行深度的数据挖掘和探索来看一下。

下面我就以获客购买流程为例，首先根据前面的数据访谈和内部沟通的情况，我们先把流程梳理一下。因为是举例，我们简化变成下图这样一个流程。你能从图中看到我们现在主要是通过购买百度关键字、直播活动和抖音的投放进行获客，然后通过电话销售的方式来进行促销，最终达成整体的购买。



这里边我们涉及以下几类数据。

百度的投放数据明细；

渠道的投放数据明细；

直播的活动数据明细；

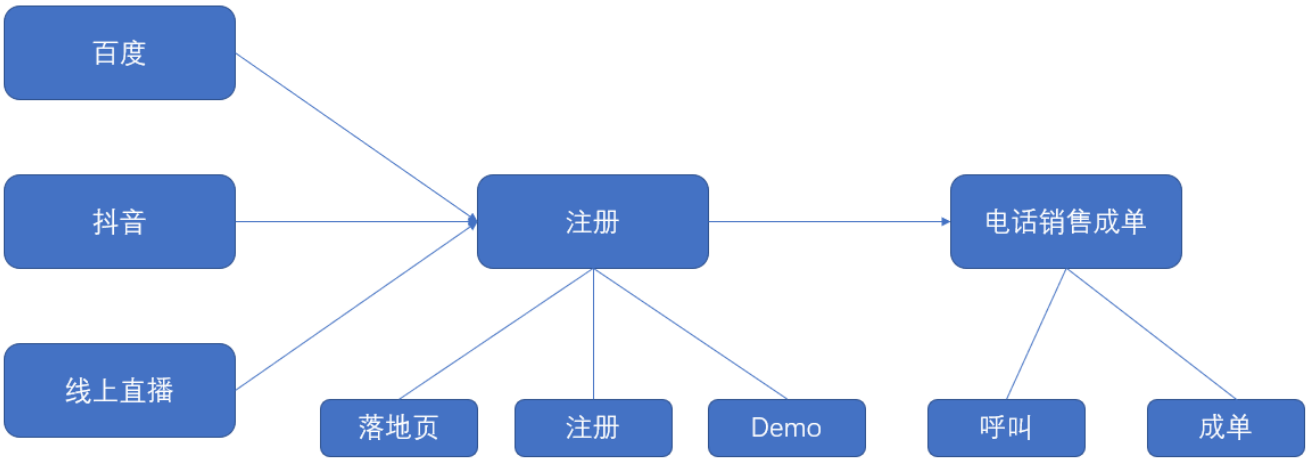
电话销售的成单数据明细。

从我们数据分析来看，这个流程其实还不完整，因为我们缺少了一些中间过程的数据，例如用户的注册数据、用户的访问网站的数据、用户打开 demo 的数据……这些数据就是我们要的采集点，我们可以把它细化为如下的数据进行分析。

用户访问明细；

用户销售数据明细。

这样我们就可以统计用户访问量和销售量，然后我们可以使用快照扩展法，继续进行细分。比如现在我们拿到某一天的相关数据，发现用户访问量还是过于粗犷了。我们应该分成落地页访问、注册页访问、demo 访问三个不同的指标。



然后每一个指标我都可以把它分到不同渠道的访问用户量和成单量，也可以针对某一个渠道再细分。比如我们可以把百度投放的数据，细分到关键字的成单数据以及落地访问量。这样我们的指标又扩展一步，变成如下的情况。

用户访问量：

落地页访问（不同渠道，不同百度关键字）；

首页访问（不同渠道，不同百度关键字）；

demo 访问（不同渠道，不同百度关键字）。

用户销售数据（不同客服，不同渠道，不同百度关键字）。

光有这些数据其实并没有解决老板提到的大方向投入成本问题，还需要过程数据。因此我们要衍生出来一些指标去和我们的最终目标进行挂钩，我们可以根据不同活动和渠道购买的量给出一个成本情况，再根据不同渠道访问量以及销售数据给出一个平均每次访问的成本价格。

我们这里面还有很多可以扩展的指标，比如销售的转化率、整个访问的留存情况、新客户的转化和老客户转化情况……最后根据我们的下节课数据讲故事的情况，可能还需要再进行

调整。

百度渠道成本（关键字）；

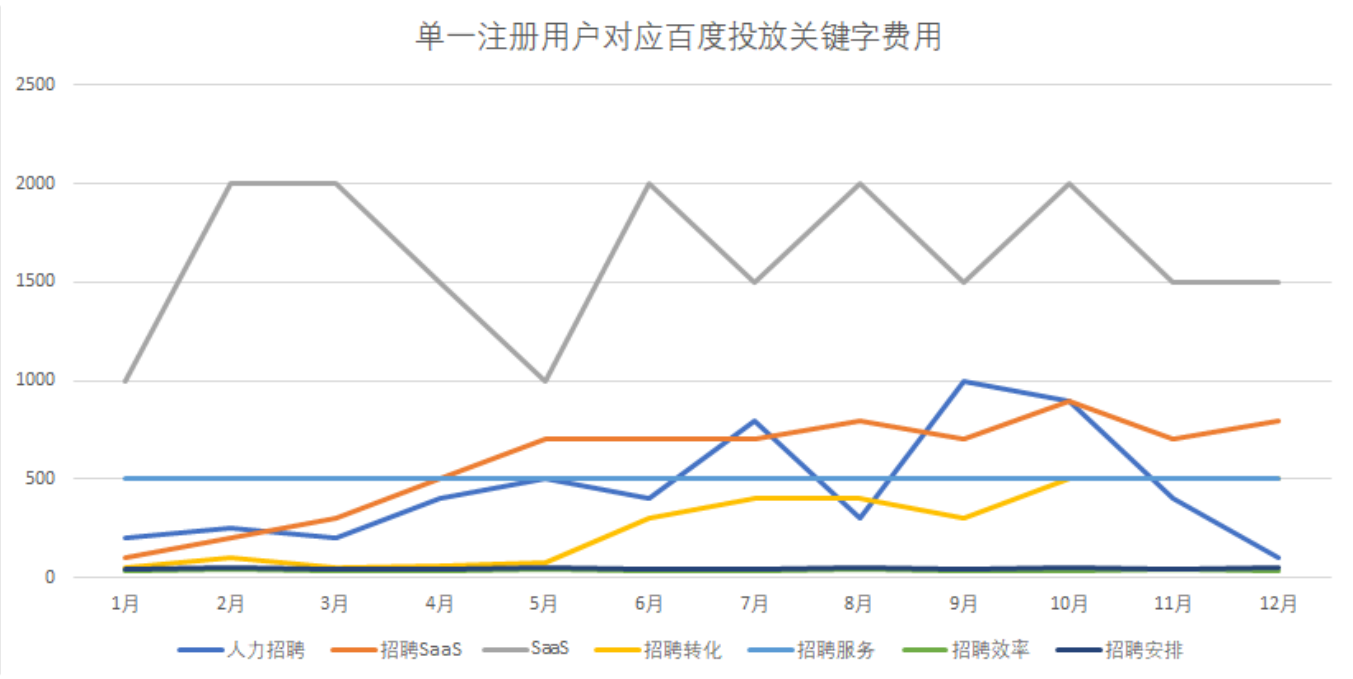
抖音广告成本；

直播活动成本；

销售转化率；

客户注册转化率。

有了这些快照扩展之后，我们拿到了更详细的一些数据指标，我们就可以使用趋势延伸法来看到这些指标在不同情况下的一些波动，比如下面的这个图。



除了这些一手数据，我们还要有一些外部的二手数据。我们可以通过圈内人士知道同行业 SaaS 的获客的成本在 3~5 万、一次活动的注册转化率应该高于 10%、成单转化率应该多于 2%……这些其实都可以作为衡量的指标。

有了这些数据，我们其实还没有完成整体的数据分析思路，因为这个思路需要我们梳理分析后，再提供新的数据反复迭代。小小预告一下，这就是我们下节课要来重点探讨的内容。

小结

小结一下，今天主要给你讲了怎么把这些数据采集起来。我们主要先从企业内部的一手数据进行相关的数据采集，这里要注意，一手数据要用明细数据直接进行数据分析，不要在企业内部还使用二手数据来混淆自己的视听。

只有企业自身的数据还不够，我们很多时候还要参考业内的二手数据来做一个衡量准绳，通过整体行业趋势给我们更多的指引。

在做数据探索的时候光有裸数据还不够，我们还要进行一些数据探索和数据延展。今天我给你介绍了三个比较常用的方法：趋势分析法、快照扩展法和衍生指标法。这几个方法可以帮助我们再从裸数据里面看到更多的延展相关数据，为我们下一步编写数据故事打好基础。

最后我举了一个具体的例子，延续前面课程的内容来给你简单剖析了一下，我们如何做深度的相关分析，这里是想带你体验一下怎么去扩展和筛选相关的重要指标。

其实我们在工作和生活当中的决策也离不开这个采集数据的过程。例如我们要决策投资、准备跳槽选工作，那我们就既要了解自己的一手数据（资金情况、学历情况、收入情况），也要拿到二手数据（市场情况、国家政策、老板和同事反馈、职位在市场中的价位等）。我们也可以做一些横向和纵向衍生指标（投资回报 / 风险比、跳槽收益率 = 跳槽成功概率 * 跳槽收益）。我们自己做决策不要只凭经验拍脑袋，收集更多的好数据才有利于我们更有效决策。

数据给你一双看透本质的眼睛，希望你可以在这个充满数据的世界里，采集到对你决策有帮助的数据，更加客观地进行决策。

课后思考

你经常采集数据的方法还有哪些？哪些外部数据资料是你经常使用的？分享出来，我们一起提高。

附录：常用的一些网站和信息来源渠道

宏观数据

经合组织开放的数据网：https://stats.oecd.org/

世界银行公开数据：https://data.worldbank.org.cn/

中国统计年鉴：http://www.stats.gov.cn/tjsj/ndsj/

统计局网站：http://www.stats.gov.cn/

新华社 - 全球经济数据:http://dc.xinhua08.com/

中国互联网络信息中心：https://www.cnnic.net.cn/hlwfzyj/hlwxzbg/

中财网：http://data.cfi.cn/

互联网数据

Alexa： https://alexa.chinaz.com/

百度指数： https://index.baidu.com/

微指数： https://data.weibo.com/index/

淘宝指数： https://shu.taobao.com/

阿里价格指数：http://topic.aliresearch.com/

Similarweb：https://www.similarweb.com/

netmarketshare：https://netmarketshare.com/

Statcounter：https://gs.statcounter.com/

行业数据库

数据汇： http://www.shujuhui.com/database/

数据圈： http://www.shujuquan.com.cn/

镝数聚：https://www.dydata.io/

联合国图书馆：http://www.oecd-ilibrary.org/

票房数据：https://www.boxofficemojo.com/charts/

中国票房数据：http://cbooo.cn/

行业分析机构：Gartner、Forrester、Bloomberg、易观、艾瑞、新榜

企业数据

巨潮资讯：http://www.cninfo.com.cn/new/index/

EDGAR：http://sec.gov

企业招股说明书、年报、半年报、季报、券商分析报告

投融投资数据

IT 桔子：http://www.itjuzi.com/

投资中国： http://www.chinaventure.com.cn/

创业邦： http://www.cyzone.cn/

36 氪： http://www.36kr.com/

分享给需要的人，Ta订阅后你可得 20 元现金奖励

赞 3 提建议

© 版权归极客邦科技所有，未经许可不得传播售卖。 页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 21 | 确定问题：与利益无关的问题都不值得数据分析和挖掘

下一篇 编辑手记 | 让生活多一些科学，少一些神学

精选留言 (3)

写留言

进化菌 神
2021-09-17

干货好多，感谢老师的分享

展开

1

范
2021-09-18

数据分析应该就是先按照结构化的形式细化分类，然后整合分析。细化的维度可以是时间维度（趋势分析法），结构的维度（类似于部门、地区、分类等，快照扩展），过程维度（细化一个商业过程从开始到结束）。

展开

那时刻 神
2021-09-17

请问老师，数据探索过程中，有您提到的三种方法。比如在探索过程中，有了一个初步的结论，如何来校对我的初步结论的正确性呢？比如有可能是数据误差或者人为后续处理数据错误导致的问题。